**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**



**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

🞠◊🞠◊🞠



**BÁO CÁO ĐỒ ÁN 1**

**ĐỀ TÀI: GOM CỤM DỮ LIỆU VỚI GIẢI THUẬT DBSCAN**

**GVHD: ThS.LÊ VĂN VINH**

**SVTH: NGUYỄN THỊ NGỌC TRÂN**

**MSSV: 16131104**

**SVTH: TRƯƠNG THÁI TUẤN**

**MSSV: 16110249**

***TP. Hồ Chí Minh, Ngày 16 tháng 12 năm 2018***

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**



**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

🞠◊🞠◊🞠



**BÁO CÁO ĐỒ ÁN 1**

**ĐỀ TÀI: GOM CỤM DỮ LIỆU VỚI GIẢI THUẬT DBSCAN**

**GVHD: ThS.LÊ VĂN VINH**

**SVTH: NGUYỄN THỊ NGỌC TRÂN**

**MSSV: 16131104**

**SVTH: TRƯƠNG THÁI TUẤN**

**MSSV: 16110249**

***TP. Hồ Chí Minh, Ngày 16 tháng 12 năm 2018***

**LỜI CẢM ƠN**

*Không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp. Nay chúng em xin được phép gửi lời cảm ơn chân thành này đến thầy Lê Văn Vinh, người đã trực tiếp hỗ trợ chúng em trong suốt quá trình định hướng chọn đề tài, hướng dẫn, nhận xét và góp ý. Nếu không có những lời hướng dẫn, những kinh nghiệm thực tiễn của thầy thì chúng em nghĩ rằng bài báo cáo đồ án một này sẽ khó có thể hoàn thiện và hoàn thành đúng thời hạn được. Một lần nữa, chúng em xin cảm ơn thầy.*

*Chúng em cũng xin chân thành cảm ơn các quý thầy cô trong khoa Công Nghệ Thông Tin đã giúp đỡ hỗ trợ kiến thức cũng như giải đáp thắc mắc của chúng em. Cùng với đó, chúng em xin được gửi cảm ơn đến các bạn cùng khóa đã cung cấp nhiều thông tin và kiến thức hữu ích giúp cho chúng em hoàn thiện đề tài hơn.*

*Bài báo cáo được thực hiện trong khoảng thời gian gần 6 tuần. Khoảng thời gian có hạn, cùng với kiến thức còn hạn chế và còn nhiều bỡ ngỡ khác do đó thiếu sót là điều không thể tránh khỏi nên chúng em rất mong nhận được những ý kiến đóng góp quý báo của các quý Thầy Cô để kiến thức của chúng em được hoàn thiện hơn sau này. Chúng em xin chân thành cảm ơn.*

*Thành phố Hồ Chí Minh, ngày 16 tháng 12 năm 2018*

*Sinh viên thực hiện 1 Sinh viên thực hiện 2*

*(ký và ghi họ tên) (ký và ghi họ tên)*

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

Giáo viên hướng dẫn

(ký và ghi họ tên)

NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

Giáo viên phản biện

(ký và ghi họ tên)

Mục lục

[I. Mô tả project 1](#_Toc532734485)

[1. Giới thiệu khái quát về gom nhóm dữ liệu 1](#_Toc532734486)

[2.Giới thiệu project 1](#_Toc532734487)

[3.Giới thiệu thuật toán 1](#_Toc532734488)

[a. Sơ đồ thuật toán 3](#_Toc532734489)

[b. Mô tả thuật toán 3](#_Toc532734490)

[c. Ứng dụng của thuật toán 5](#_Toc532734491)

[d. Ưu nhược điểm của thuật toán 5](#_Toc532734492)

[II. Quá trình thực hiện 6](#_Toc532734493)

[1.Mẫu thiết kế code 6](#_Toc532734494)

[1.1. Bảng danh mục các lớp được sử dụng trong chương trình 6](#_Toc532734495)

[1.2. Bảng mô tả các phương thức trong lớp Point 7](#_Toc532734496)

[2.Demo với dữ liệu giả lập 8](#_Toc532734497)

[3.Ứng dụng của thuật toán DBScan 11](#_Toc532734498)

[3.1.Phân tích bài toán 11](#_Toc532734499)

[4.Phân công công việc 17](#_Toc532734500)

[III. Kết luận 18](#_Toc532734501)

[1.Đánh giá project 18](#_Toc532734502)

[2.Vấn đề khi thực hiện project 18](#_Toc532734503)

[3.Ưu điểm và nhược điểm của project 18](#_Toc532734504)

[a. Ưu điểm 18](#_Toc532734505)

[b. Khuyết điểm 18](#_Toc532734506)

[IV. Tài liệu tham khảo 18](#_Toc532734507)

Danh mục các hình

[Hình I.1 3](#_Toc532734112)

[Hình I.2 4](#_Toc532734113)

[Hình II.1 10](#_Toc532734114)

[Hình II.2 11](#_Toc532734115)

[Hình II.3 12](#_Toc532734116)

[Hình II.4 13](#_Toc532734117)

Danh mục các bảng

[Bảng II.1 7](#_Toc532734414)

[Bảng II.2 8](#_Toc532734415)

[Bảng II.3 9](#_Toc532734416)

# Mô tả project

## Giới thiệu khái quát về gom nhóm dữ liệu

Gom nhóm dữ liệu là quá trình tiền xử lý cho phân lớp dữ liệu, tập hợp các đối tượng lại thành những nhóm có ý nghĩa. Trong quá trình gom nhóm này, kết quả chúng ta sẽ nhận được các nhóm mà các đối tượng trong cùng một nhóm có những tính chất chung và có những tính chất khác với các đối tượng ở nhóm khác.

Cho CSDL D={t1,t2,…,tn) và số nguyên k, gom nhóm là bài toán xác định ánh xạ f: D🡪{1,…,k) sao cho mỗi ti được gán vào một nhóm Kj, 1≤ j ≤ k.

Không giống các bài toán phân lớp, các nhóm không được biết trước. Được ứng dụng trong nhiều lĩnh vực như nhận dạng vân tay, móng mắt; xử lý ảnh; đặc biệt trong tiếp thị, nghiên cứu sở thích, xu hướng của khách hàng; gom nhóm các tài liệu liên quan dễ dàng cho việc tìm kiếm,…

## Giới thiệu project

Đồ án sử dụng ngôn ngữ C# và ứng dụng thuật toán DBScan để gom nhóm dữ liệu trong hình học không phẳng. Cho một tập dữ liệu có sẵn, nhờ vào thuật toán DBScan, chúng ta sẽ gom nhóm được các cụm dữ liệu có tính chất giống nhau dựa theo Eps và Minpts.

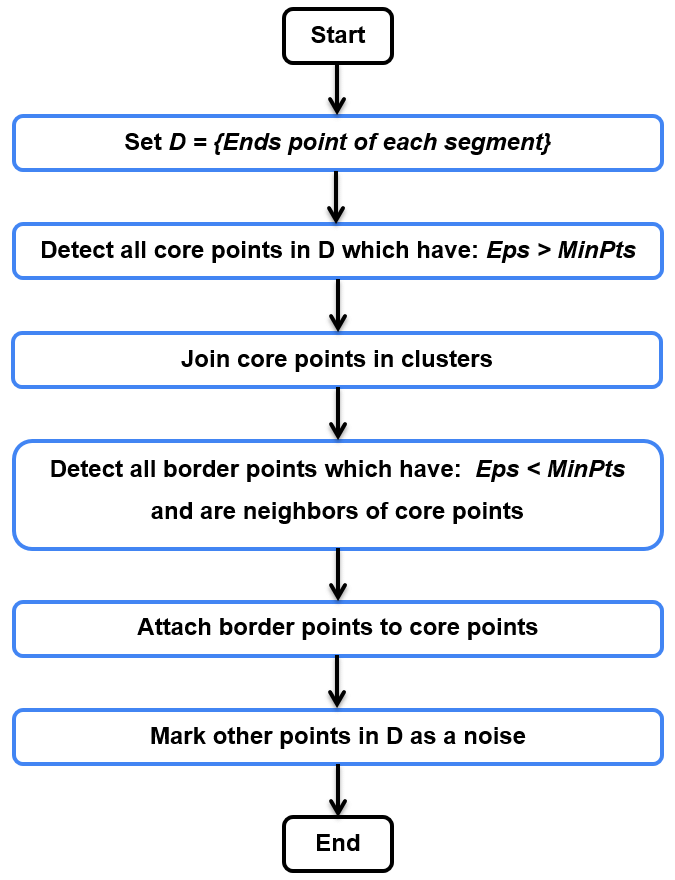
## Giới thiệu thuật toán

Thuật toán DBScan

* DBScan (Density- based spatial clustering of applications with noise) là giải thuật gom cụm hiệu quả dựa trên mật độ (minPts) do Martin Ester và các tác giả khác đề xuất, cho phép phát hiện các cụm có hình dạng bất kỳ, làm việc tốt với dữ liệu nhiễu trong không gian vì có khả năng phát hiện nhiễu tốt, hiệu quả với cơ sở dữ liệu lớn.
* Thuật toán gồm 2 tham số (parameter) là eps và minPts. Từ 1 nút chưa được chọn, kiểm tra các điểm gần nhất, nếu số lượng các điểm này lớn hơn giá trị minPts thì bắt đầu 1 nhóm mới. Nếu không sẽ đánh dấu là điểm nhiễu. Điểm nhiễu này vẫn có thể thuộc 1 nhóm khác, khi đó sẽ bỏ đánh dấu điểm nhiễu. Cứ thế mở rộng ra đến khi không thể tìm thêm điểm mới cho nhóm.
* Trong thực hiện, giải thuật cần bước tìm kiếm k láng giềng, nên DBSCAN có độ phức tạp bậc hai theo số lượng phần tử.

Độ phức tạp : tính toán (N\*logN), dữ liệu (N^2).

### Sơ đồ thuật toán

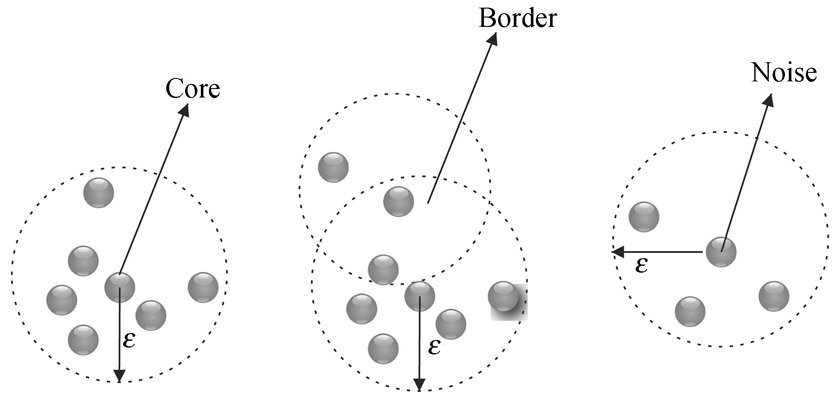


*Hình I.1*

### Mô tả thuật toán

Một số khái niệm:

* Density (mật độ): số lượng các điểm trong một bán kính cho trước (eps).
* Vùng lân cận của điểm p: là tập hợp các đối tượng q sao cho khoảng cách từ điểm q đến điểm p nhỏ hơn eps.
* minPts: số lượng điểm tối thiểu được cho trước.
* Core point (điểm nòng cốt): Một điểm p được gọi là core point nếu số lượng các đối tượng có mật độ có thể đạt được từ p ít nhất bằng minPts.
* Border point (điểm biên): Một điểm p được gọi là border point nếu số lượng các đối tượng có mật độ có thể đạt được từ p ít hơn số lượng minPts cần đạt được và có ít nhất 1 core point trong khu vực lân cận của điểm p.
* Noise point (điểm nhiễu): là đối tượng không thuộc core point hoặc border point và không thuộc về bất kì nhóm nào.



*Hình I.2*

Các bước thực hiện của thuật toán được mô tả như sau:

1. Chọn ngẫu nhiên một đối tượng p (lúc này p gọi là hạt giống). Nếu p là đối tượng nhiễu (noise point) hay cá biệt thì chúng ta sẽ không xét.
2. Tìm tất cả các đối tượng có mật độ có thể đạt được từ p theo eps và minPts (đi theo một chiều và vào nhiều hướng).
3. Nếu p là đối tượng nòng cốt (core point) thì hình thành nhóm. Khi tìm tất cả các đối tượng có mật độ có thể đạt được từ p, thực tế là chúng ta đã tìm các đối tượng có mật độ liên thông lớn nhất, bởi vì quá trình này là quá trình lặp đi lặp lại.

Nếu p không phải là đối tượng nòng cốt (core point) mà là đối tượng biên( border point) thì thuật toán sẽ xem xét đối tượng tiếp theo trong cơ sở dữ liệu.

1. Thuật toán dừng khi tất cả các đối tượng đã được xử lý.

### Ứng dụng của thuật toán

* Scientific literature: Tài liệu khoa học, đặc biệt trong lĩnh vực nghiên cứu tiếp thị, giúp nghiên cứu các sở thích, xu hướng của khách hàng.
* Images of satellite: Xử lý ảnh, phân tích dữ liệu không gian lấy từ các vệ tinh.
* Crystallography of x-ray.
* Anomoly detection in temperative data: Phát hiện các bất thường trong dữ liệu, ví dụ xâm nhập mạng trái phép hoặc sử dụng thẻ tín dụng gian lận.

### Ưu nhược điểm của thuật toán

* Ưu điểm
* Không cần xác định trước số lượng các nhóm. Điều này trái ngược với K-Means.
* DBSCAN có thể tìm được những nhóm có dạng hình học đặc biệt như một nhóm bị bao kín hoàn toàn bởi 1 nhóm khác. Với tham số minPts, DBSCAN giảm được hiệu ứng Single\_link (các nhóm khác nhau bị liên kết với nhau bởi 1 đường mỏng).
* DBSCAN có khái niệm về nút (mẫu) nhiễu.
* DBSCAN cần 2 tham số, và không nhạy cảm với sự thay đổi thứ tự trong dữ liệu.
* Nhược điểm
* Kết quả phụ thuộc rất lớn vào cách tính khoảng cách giữa các điểm

( cách lựa chọn eps và Minpts).

* DBSCAN không hiệu quả khi phân nhóm 1 tập hợp dữ liệu có mật độ phân tán khác nhau.
* Từ một đối tượng phải tính tất cả các đối tượng có mật độ đạt được với nó, rồi từ đối tượng có mật độ đạt được tới nó phải tính tiếp tục, cứ liên thông như vậy nên độ phức tạp của thuật toán rất cao.

# Quá trình thực hiện

## Mẫu thiết kế code

### Bảng danh mục các lớp được sử dụng trong chương trình

Sinh viên phụ trách: Trương Thái Tuấn

Nguồn tham khảo: <https://www.c-sharpcorner.com/uploadfile/b942f9/implementing-the-dbscan-algorithm-using-C-Sharp/>

*Bảng II.1*

|  |  |  |
| --- | --- | --- |
| STT | Tên lớp | Mục đích chính của lớp trong chương trình |
| 1 | class Point | Khởi tạo các thuộc tính của điểm, tính khoảng cách giữa 2 điểm |
| 2 | static class DBSCAN | Khởi tạo dữ liệu, các điểm cần gom cụm, bán kính erp, số điểm tối thiểu minPts, gom cụm |
| 3 | static List<List<Point>> TaoCum(List<Point> points, double eps, int minPts) | Tạo cụm bằng các điểm có sẵn |
| 4 | static List<Point> TimDiemLanCan(List<Point> points, Point p, double eps) | Tìm các điểm lân cận |
| 5 | static bool MoRongCum(List<Point> points, Point p, int CumId, double eps, int minPts) | Mở rộng cụm, và xét hết cơ sở dữ liệu |

### Bảng mô tả các phương thức trong lớp Point

Sinh viên phụ trách: Trương Thái Tuấn

Nguồn tham khảo: <https://www.c-sharpcorner.com/uploadfile/b942f9/implementing-the-dbscan-algorithm-using-C-Sharp/>

*Bảng II.2*

|  |  |  |  |
| --- | --- | --- | --- |
| TT | Phương thức | Mục đích | Tên file, số thứ tự dòng chứa khai báo |
| 1 | public Point(int x, int y) | Khởi tạo tọa độ của 1 điểm (x,y) | Dòng 10 |
| 2 | public override string ToString() | Trả về biểu diễn chuỗi | Dòng 15 |
| 3 | public static int khoangcach(Point p1, Point p2) | Tính khoảng cách giữa 2 điểm p,q | Dòng 19 |

## Demo với dữ liệu giả lập

Các parameter cần nhập

* points: các điểm cần gom cụm
* eps: bán kính
* minPts: số lượng điểm tối thiểu
* time: Thời gian thực hiện chương trình.

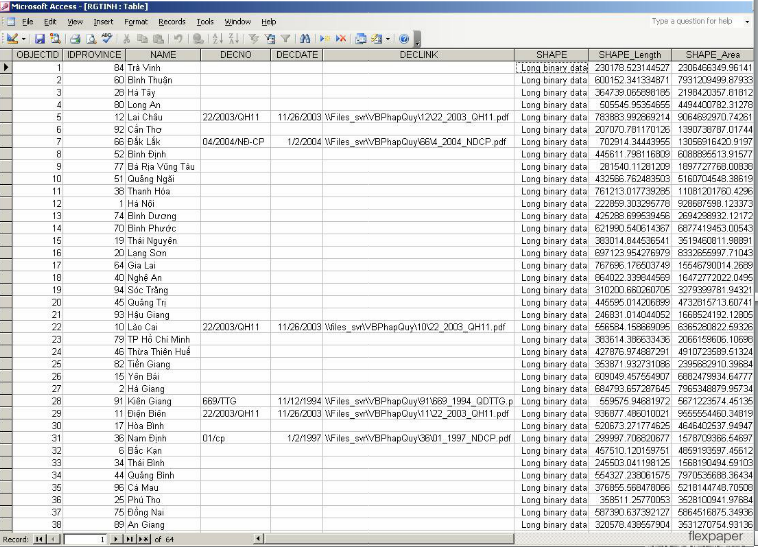
*Bảng II.3*

|  |  |
| --- | --- |
| Dữ liệu nhập | Kết quả |
| points.Add(new Point(0, 100));  points.Add(new Point(0, 200));  points.Add(new Point(0, 275));  points.Add(new Point(100, 150));  points.Add(new Point(200, 100));  points.Add(new Point(250, 200));  points.Add(new Point(0, 300));  points.Add(new Point(100, 200));  points.Add(new Point(600, 700));  points.Add(new Point(650, 700));  points.Add(new Point(675, 700));  points.Add(new Point(675, 710));  points.Add(new Point(675, 720));  points.Add(new Point(50, 400));  points.Add(new Point(20, 200));  points.Add(new Point(280, 80));  points.Add(new Point(300, 100));  points.Add(new Point(400, 100));  points.Add(new Point(265, 85));  points.Add(new Point(350, 150));  double eps = 100.0;  int minPts = 3; |  |

## Ứng dụng của thuật toán DBScan

### a. Cơ sở dữ liệu

Bản đồ Đơn vị hành chính Việt Nam dưới dạng dữ liệu thuộc tính, được cung cấp từ Cục Công nghệ thông tin - Bộ Tài nguyên và Môi trường.



*Hình II.1*

### b. Công nghệ sử dụng

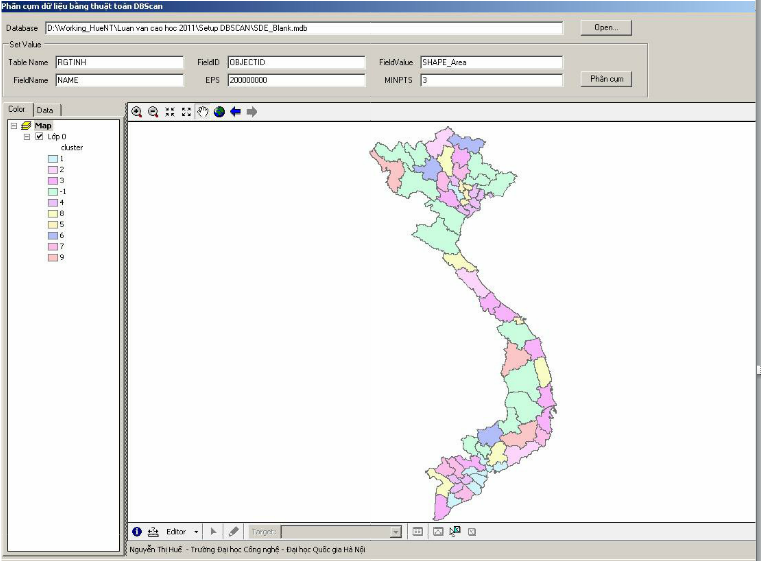
Sử dụng thư viện ArcGIS Engine và công nghệ lưu trữ dữ liệu không gian Geodatabase của ESRI.

### c. Ý tưởng

Những đối tượng trong cùng một cụm phải có mật độ các điểm lân cận vượt qua một ngưỡng nào đó. Ví dụ như một điểm chỉ có thể thuộc một cụm khi có không ít hơn 3 đối tượng trong lân cận Eps.

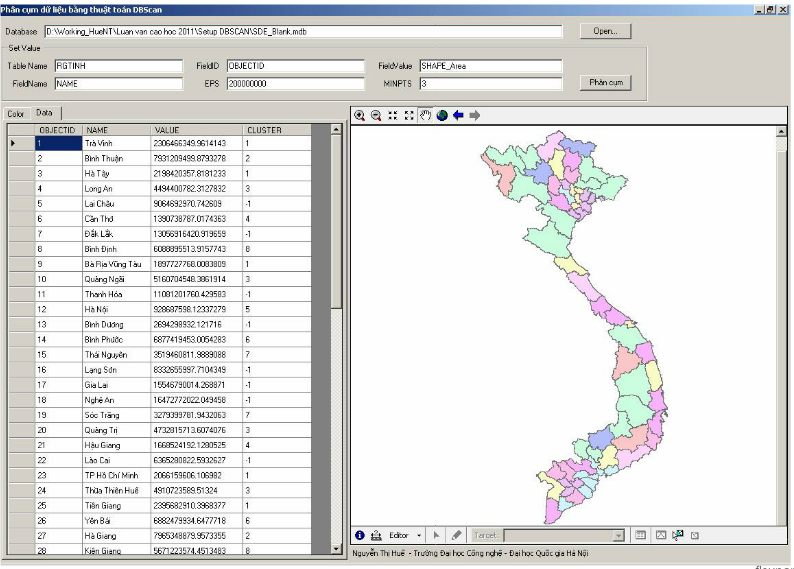
### d. Kết quả ứng dụng

Với Minpt = 3, Epxilon=200000000🡪Dữ liệu bản đồ được chia làm 9 cụm và một cum dữ liệu nhiễu (có giá trị là -1) tương ứng với 10 màu thể hiện trên bản đồ.



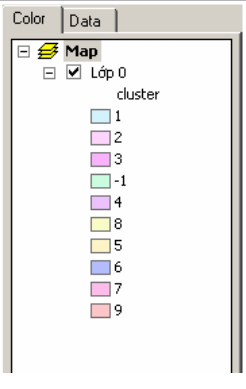
*Hình II.2*

Kết quả dữ liệu phân cụm được thể hiện tương ứng với màu trên bản đồ



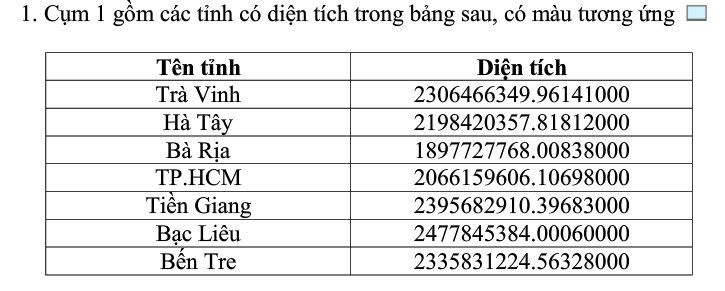
*Hình II.3*

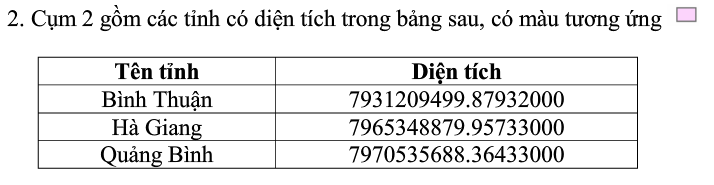
Các cụm tương ứng với màu thể hiện trên bản đồ

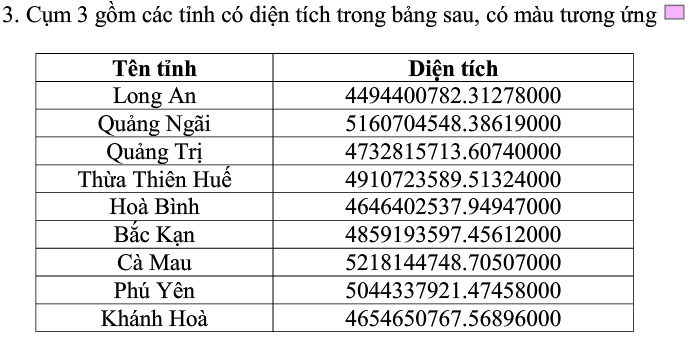


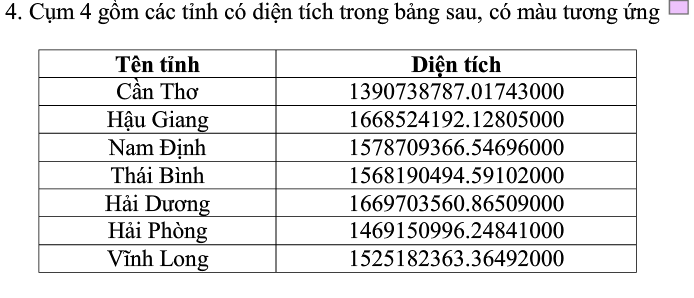
*Hình II.4*

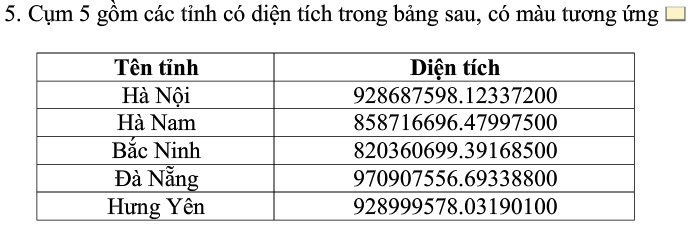
Ứng với mỗi màu là giá trị ngưỡng của từng vùng thể hiện diện tích của vùng đó:

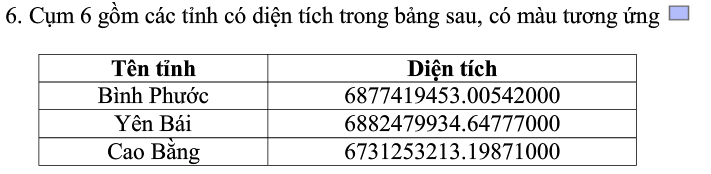


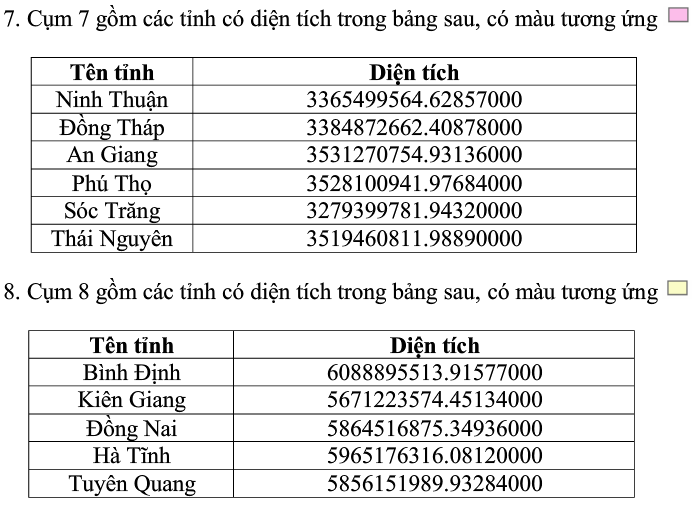


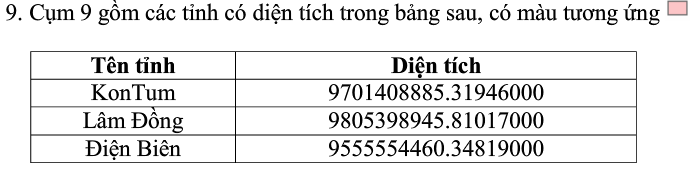


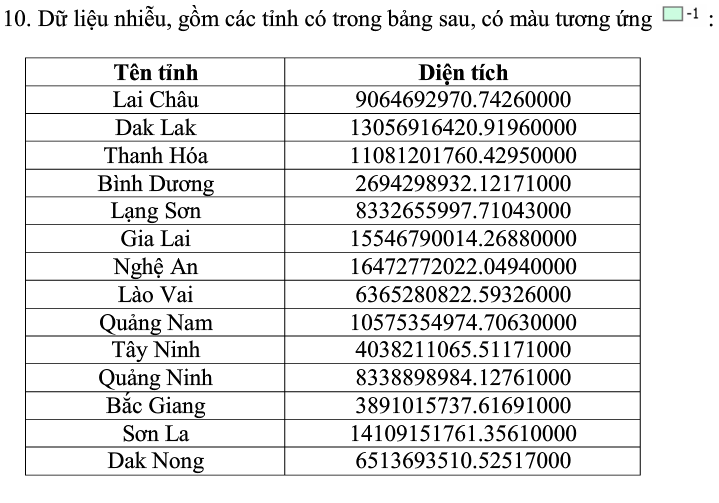












Kết luận:

Trên đây là cài đặt thử nghiệm giải thuật DBScan với bài toán phân cụm dữ liệu bản đồ cả nước dựa vào diện tích các tỉnh. Đây là một bài toán đơn giản, là một ứng dụng phân cụm dựa trên diện tích (hoặc tương tự với các tham số khác như: phân cụm về dân số, loại đất, động đất, dịch bệnh,…).

## Phân công công việc

|  |  |  |
| --- | --- | --- |
| Tên SV | Đánh giá chung phần trăm đóng góp | Mô tả khái quát mảng công việc SV thực hiện trong đồ án. |
| Nguyễn Thị Ngọc Trân | 40% | Viết báo cáo |
| Trương Thái Tuấn | 60% | Viết code |

Bảng 10. Phân công công việc

# Kết luận

## Đánh giá project

* Project thực hiện được đầy đủ yêu cầu đề ra, ứng dụng thuật toán gom cụm để gom cụm dữ liệu vào một nhóm cần thiết.

## Vấn đề khi thực hiện project

* Việc tìm kiếm dữ liệu mất nhiều thời gian.
* Dữ liệu về thuật toán còn nhiều hạn chế, không có đủ thời gian làm đề tài.

## Ưu điểm và nhược điểm của project

### Ưu điểm

* Hiểu được cách thức hoạt động một cách rõ ràng, chạy được code minh hoạ đơn giản của DBScan.

### Khuyết điểm

* Cài đặt thuật toán còn khá đơn giản.
* Chưa làm được các bài toán ứng dụng vào thực tế.

# Tài liệu tham khảo

[1].http://www.sthda.com/english/wiki/print.php?id=246

[2].<https://cs.wmich.edu/alfuqaha/summer14/cs6530/lectures/ClusteringAnalysis.pdf>

[3].<http://www.cit.ctu.edu.vn/~dtnghi/rech/dir_0/Duy-et-al-New.pdf>

[4].http://www.thuvientailieu.vn/tai-lieu/luan-van-nghien-cuu-cac-ky-thuat-phan-cum-du-lieu-va-ung-dung-4144/